心理学报 2022, Vol. 54, No. 10, 1277−1292
Acta Psychologica Sinica

https://doi.org/10.3724/SP.J.1041.2022.01277

基于作答时间数据的改变点分析在检测加速作答中的探索——已知和未知项目参数*

钟小缘[1] 喻晓锋[1] 苗 莹[1] 秦春影[2] 彭亚风[1] 童 昊[1]
([1]江西师范大学心理学院，南昌 330022)
([2]南昌师范学院数学与信息科学学院，南昌 330032)

摘 要 相对于传统的离散作答数据，作答时间作为连续数据，可以提供更多信息。改变点分析(change point analysis)技术在心理和教育领域是一个比较新的技术。本文一方面对改变点分析在心理测量领域的应用进行了一个综合的总结和分析；另一方面，将基于作答数据的两种改变点分析统计量推广到作答时间数据，将改变点分析技术应用到测验异常作答模式：加速作答 speededness 的检测上。采用两种检验方法：似然比检验和 Wald 检验，分别在已知和未知项目参数的条件下，实现异常作答模式的检测。结果表明，所采用的方法对于加速作答行为的检测具有很高的检验力，同时能够很好的控制 I 类错误率。实证数据分析进一步表明本文中所使用的方法具有应用价值。
关键词 改变点分析法，异常作答行为，作答时间，加速作答，统计过程控制
分类号 B841

1 引言

高质量的测量数据是对考生能力做出精确评估的保障。然而实际情况中会存在许多导致系统误差的因素对数据的质量产生影响，其中最常见的因素是考生的异常作答行为。测验过程中常见的异常作答行为有热身效应、加速作答等等(骆方 等, 2020; 张龙飞 等, 2020)。考生出现异常作答行为后，其作答数据与其正常作答时的数据有着显著的不同。考生出现异常作答行为时产生的数据称为异常作答数据或异常作答模式。测验数据中包含异常作答数据会降低其自身及整体测验数据的质量，从而对后续的分析结果产生一系列的不良影响，例如造成模型与数据的不拟合、被试与题目参数估计的偏差(Stefan et al,. 2016)，影响考试的信度和效度(Guo et al,. 2009)等等。因此，检测测验中的异常作答数据是非常重要和关键的，研究者们也一直在寻找相关的解决方法(如 Bejar, 1985; Evans & Reilly, 1972; Shao et al,. 2016; Bradlow et al,. 1998; McLeod et al,. 2003; Wise & Kong, 2005; Yu & Cheng, 2019, 2022)。

改变点分析法(change point analysis, CPA; Page, 1955; Shao et al., 2016; Sinharay, 2016)是检测过程数据中存在异常的常用方法，多用于判断序列数据中是否存在分布形态的变化，即数据中是否存在改变点。它的基本原理是：在一组序列数据中，样本按照时间先后顺序排列，在不同的时间点，样本有不同的取值。如果从某个时刻开始，形成时间序列的样本不再服从原来的分布或者样本特征，比如均值或方差等发生了显著的变化，即表明数据中出现了改变点(Hawkins et al., 2003)，改变点的出现说明事物发生了质的变化。

近年来研究者们将 CPA 引入心理与教育测量领域检测测验中的异常作答行为或异常作答模式(Zhang, 2014; Shao, 2016; Shao et al., 2016; Sinharay,

收稿日期: 2021-08-25
* 全国教育科学规划项目(BGA210060)；江西省社会科学基金项目(21JY06)；国家教育部考试中心科研规划课题(GJK2021025)；江西省高校人文社会科学项目(XL20202)；南昌市教育大数据智能技术重点实验室(2020-NCZDSY-012)；江西省教育厅科技项目(GJJ191691, GJJ191128)资助。
通信作者: 喻晓锋, E-mail: xyu6@jxnu.edu.cn

2016, 2017a, 2017b, 2017c; Yu & Cheng, 2019, 2022)。Shao 等(2016), Sinharay (2016), Yu 和 Cheng (2019, 2022)等研究显示了 CPA 在检测异常作答行为或异常作答模式上的优势。在测验过程中，考生在每道题目上的作答形成了独特的序列数据。一般情况下考生的作答数据会服从某种分布，例如考生的作答时间数据通常会服从于对数正态分布。当考生出现异常作答行为后，由于考生的作答行为不同于正常作答时的行为，因此考生的作答数据也会发生性质上的改变。异常作答行为发生的位置(测验中题目的编号)即为数据发生变化的点。

使用 CPA 检测测验中的异常作答行为或者异常作答模式，可以从两种数据入手，一种是考生的作答数据(response)，另一种是作答时间数据，即考生作答每道题所用的时间(response time)。作答时间数据是一种连续数据，同时包含了考生能力信息和题目信息(Marianti et al,.2014)，对于提高考生能力估计的精度与优化测验设计有很大的帮助；如今随着新技术的发展，计算机测验与在线评估越来越多，作答时间数据的获取也变得更加便利，逐渐获得学者们的关注。例如 van der Linden 和 van Krimpen-Stoop (2003)使用作答时间数据检测考生预知试题以及加速作答; van der Linden 和 Guo (2008), Pan 和 Wollack (2021)等使用作答时间数据检测测验中考生预知试题的情况等等。还有研究者基于不同的应用场景构建作答时间模型，结果显示引入作答时间数据有助于模型的参数估计等，拓宽了作答时间数据的使用范围(Wang & Xu , 2015; 郭小军，罗照盛, 2019; 詹沛达, 2019; 詹沛达 等, 2020)。

以往基于 CPA 检测测验中异常作答行为的研究多是基于作答数据，如今作答时间数据的优势已经凸显，在数据分析中引入或结合作答时间数据是非常重要的发展趋势。另外，由于加速作答是众多异常作答行为中最常见和普遍的(Goegebeur et al,. 2008)，对于测验数据质量有非常大的负面影响，受到很多研究者的关注(比如 Bolt et al,. 2002; Oshima, 1994; Suh, et al,.2012; Yu & Cheng, 2022 等)。因此本研究拟聚焦于作答时间数据，在已知项目参数和未知项目参数条件下，分别使用 CPA 方法检测由加速作答行为造成的异常作答模式。需要注意的是，CPA 方法本质上是检测异常数据的方法，因此它同样可用于检测由其他异常作答行为如题目预知，热身效应等造成的异常作答模式。下面首先介绍 CPA 技术。

2 改变点分析 CPA 技术

CPA 广泛应用于生物学、统计学和经济学领域，虽然已有学者将它引入教育与心理测量领域，但它还没有得到很好的开发。基于 CPA 检测异常作答行为的已有研究主要有：Zhang (2014), Shao 等(2016), Shao (2016), Sinharay (2016, 2017a, 2017b, 2017c), Yu 和 Cheng (2019, 2022)。其中，Zhang (2014)关注的是考生预知试题信息造成的试题泄露现象，并提出了一种实时序列试题监控方法。

Shao 等(2016)基于 CPA 使用似然比检验探测加速作答行为，这种方法不仅可以将考生分为加速组和非加速组，还能比较准确地找到考生开始出现异常作答行为的位置。异常行为发生的位置使得测验管理人员通过去除可疑的加速反应来提高能力估计的精确性，并且为实际测验中设置合适的测验长度提供参考。Shao 等(2016)使用检验统计量为

$$\Delta l_i = 2(l_i^{H_a} - l_i^{H_0}) \tag{1}$$

上面的公式中，$l_i^{H_a}$ 和 $l_i^{H_0}$ 分别表示考生出现加速作答行为和正常作答时的对数似然值。当给定考生 i 的得分数据，可以使用 MLE (Baker & Kim, 2004)等算法估计出考生的能力 θ_i，进一步得到 $l_i^{H_0}$。而 $l_i^{H_a} = l_i^{j-} + l_i^{j+}$，$l_i^{j-}$ 和 l_i^{j+} 分别表示基于 j 为分界点的两个子测验(第 1 个子测验包含题目 1，…，题目 j；第 2 个子测验包含题目 $j+1,\cdots$，题目 n)所对应的似然函数。Δl_i 达到最大值的位置即考生开始加速作答的位置，它的零分布与临界值可以通过置换分布(Shao et al., 2016)、经验分布(Yu & Cheng, 2022)或理论近似分布(Sinharay, 2016)获得。

Sinharay (2016)使用 3 种 CPA 统计量考察 CAT 中的被试拟合，并计算 3 种统计量的 I 类错误率和检验力。研究中使用近似零分布检验 CPA 统计量。研究结果显示 CPA 统计量在检测包含异常作答行为的考生方面具有良好的性能。下面是 3 个 CPA 统计量：

$$W_j = \frac{(\hat{\theta}_{1j} - \hat{\theta}_{2j})^2}{\dfrac{1}{I_{1j}(\hat{\theta}_0)} + \dfrac{1}{I_{2j}(\hat{\theta}_0)}} \tag{2}$$

$$\begin{aligned} L_j = -2\{&L(\hat{\theta}_0; Y_1, Y_2, \cdots, Y_n) - \\ &L(\hat{\theta}_{1j}; Y_1, Y_2, \cdots, Y_j) - \\ &L(\hat{\theta}_{2j}; Y_{j+1}, Y_2, \cdots, Y_n)\} \end{aligned} \tag{3}$$

$$S_j = \frac{(\nabla(\hat{\theta}_0; Y_1, Y_2, \cdots, Y_j))^2}{I_{1j}(\hat{\theta}_0)} +$$

$$\frac{(\nabla(\hat{\theta}_0;Y_{j+1},Y_{j+2},\cdots,Y_n))^2}{I_{2j}(\hat{\theta}_0)} \quad (4)$$

由于变化点未知, 3 个检验统计量如下:

$$W_{\max} = \max_{1\leqslant j\leqslant n-1} W_j \quad (5)$$

$$L_{\max} = \max_{1\leqslant j\leqslant n-1} L_j \quad (6)$$

$$S_{\max} = \max_{1\leqslant j\leqslant n-1} S_j \quad (7)$$

$I_{1j}(\hat{\theta}_0)$ 和 $I_{2j}(\hat{\theta}_0)$ 是当 $\theta=\hat{\theta}_0$ 时, 分别基于试题 1 至试题 j 和试题 j+1 至试题 n 估计的 Fisher 信息量。$\nabla(\hat{\theta}_0;Y_1,Y_2,\cdots,Y_j)$ 是当 $\theta=\hat{\theta}_0$ 时 $Y_1,Y_2,\cdots,Y_j$ 对数似然的一阶导函数。此外, Sinharay (2016)还使用 ROC (receiver operating characteristics)曲线比较 CUSUM 程序和 CPA 方法的表现。研究结果表明, 在很多条件下基于 CPA 的方法比基于 CUSUM 的方法更占优势。

Sinharay (2017a)提出了基于似然比检验(L)和拉格朗日乘数检验(也称得分检验 score test)的统计量(R)来探测考生预知试题信息的异常行为。Sinharay (2017a)将测验分为两部分, 分别为 s 和 $\bar{s}$, 正常考生(没有从预知试题信息中获益)基于 s 和 $\bar{s}$ 的得分的后验分布以及能力估计值是非常接近的, 而当考生预知某些试题的信息时, 这两部分的后验分布以及能力估计值应该具有显著差异。

$$L = 2[L(\hat{\theta}_s;y_j,j\in s)+L(\hat{\theta}_{\bar{s}};y_j,j\in\bar{s})- L(\hat{\theta}_0;y_j,j=1,2,\cdots,n)] \quad (8)$$

$$R=\frac{[\nabla(\hat{\theta}_0;y_j,j\in s)]^2}{I_s(\hat{\theta}_0)}+\frac{[\nabla(\hat{\theta}_0;y_j,j\in\bar{s})]^2}{I_{\bar{s}}(\hat{\theta}_0)} \quad (9)$$

与上一研究类似, $\hat{\theta}_s,\hat{\theta}_{\bar{s}},\hat{\theta}_0$ 分别为基于 $s,\bar{s}$ 和所用试题 $(j=1,2,\cdots,n)$ 的能力估计值, y_i 表示得分, $L(\hat{\theta}_s;y_j,j\in s)$ 为 s 部分试题得分的对数似然, $\nabla(\hat{\theta}_0;y_j,j\in s)$ 为对数似然的一阶导函数(Baker & Kim, 2004), $I_s(\hat{\theta}_0)$ 表示能力为 θ_0 时 s 部分的题目信息量之和。其余类似。研究表明, 这两种统计量可用于适应性与非适应性测验, 二级和多级计分题目, 且服从渐近的标准正态分布, 这对实际应用非常有利。另外研究结果显示, 新的统计量具有可控的 I 类错误率和相对较高的检验力。

Sinharay (2017b)提供了 CPA 检测的一般过程, 对如何选择合适的统计量、相应临界值的获取方法和一些有关问题进行了讨论, 提出了解决方法, 并基于 Rasch 模型, 通过 3 个真实数据的例子说明了如何在心理测量问题中应用 CPA 检测做出重要的推论。

基于三参数 Logistic 模型(3PLM; Birnbaum, 1968), Sinharay (2017c)比较了两种 CPA 统计量, 一种是基于似然比检验的统计量 L_s, 另一种是后验偏移统计量 PSS (Belov, 2016), 在检测考生存在预知试题信息行为上的表现。这两个统计量的检测原理与 Sinharay (2017a)类似, 当考生从预知试题中获益时, 他/她在测验 c 部分(包含泄露试题的部分)上的能力估计值或后验分布与测验 u 部分(不包含泄露试题的部分)上的能力估计值或后验分布距离较远, 且 c 部分能力值高于 u 部分或者 c 部分后验分布在 u 部分后验分布的右边。L_s 的绝对值等于 L (Sinharay, 2017a)的平方根。后验偏移统计量量化了两种后验分布的距离。结果显示两种统计量的一类错误率和探测率非常接近。

Yu 和 Cheng (2019)提出了一种基于加权残差的 CPA 统计量, 并与其他 3 种 CPA 统计量(Sinharay, 2016)比较了在探测后程随机作答(back random responding, BRR)行为上的表现。结果显示, 基于加权残差的 CPA 统计量可以在 20 个题目及以上的测验中较准确的检测出 BRR。和其他 3 种统计量相比, I 类错误率都能很好的控制, 检验力高出 17%~42%。实证研究的结果也显示了基于加权残差的 CPA 统计量在检测 BRR 上的实用性。Yu 和 Cheng (2022)比较了 12 种 CUSUM 统计量与 3 种 CPA 统计量在检测加速作答行为上的性能。为检测这两类方法的稳健性与灵活性, 研究中考虑了两种不同的加速机制即速度渐变与速度突变, 即能力渐变模型(graduate change model, GCM)和能力突变的混合模型(hybrid model, HM)来模拟。除测验长度外, 还考虑了加速行为的流行程度(有加速行为的考生在考生总体中所占的比例), 严重程度(出现加速行为的考生在测验中受加速行为影响的题目比例)等变量。

为了对基于 CPA 的方法在心理测量中的应用有一个更具体的了解, 我们提供了如下的表 1。表 1 中详细列出了有关的研究, 并且从不同的角度对这些研究进行了一个综合的总结, 从其中我们可以得出很多有用的信息。例如, 从各研究基于的数据来源看, 所有研究都是基于作答数据; 从临界值的获取方式看, 只有两项研究特殊, 分别是 Shao 等(2016)是基于置换分布获得临界值和 Sinharay (2016)是采用近似临界值, 其余使用的都是经验临界值。采用经验临界值的好处是它实现简单, 适用于所有的统计量, 不像近似临界值只适用于部分统计量, 也不像使用置换分布那样需要相当长的时间

表 1 与心理和教育测量有关的 CPA 研究

文献	测验类型	研究类型	数据类型	测验长度	题目计分	样本量	模型	统计量	临界值
Zhang (2014)	A	S&E	R	40	2	10,000	3PLM	$\hat{Z}_{nm}$	经验临界值
Shao, Li & Cheng (2016)	NA	S&E	R	50,80(S)32(E)	2	500 (S), 5000 (E)	2PLM	∇l_i	基于置换分布得到临界值
Shao (2016)	NA&A	S&E	R	50(A),40,50, 60, 80(NA)	2	1000 (NA &A)	Rasch &2PLM	$\nabla l_i, W_i^{(j)}$	经验临界值
Sinharay (2016)	A	S&E	R	20,40,60,100(S) [60-250] (E)	2	100,000 (S) 70,000 (E)	Rasch	$W_{\max}, L_{\max}, S_{\max}$	近似临界值
Sinharay (2017a)	NA	S&E	R	170(E,NA),170 (S, NA),50 (S,A)	2	1636, 1644 (E, NA), 100,000 (S,NA), 1000 (S,A)	Rasch (NA), 3PLM (A)	L_S, R_S	经验临界值
Sinharay (2017b)	A	S&E	R	[60,250],170	2	70,000, 1644	Rasch	$T_{\max}, L_{\max}, W_{\max}$	经验临界值
Sinharay (2017c)	NA&A	S&E	R	170 (E,NA),100 (S, NA),50 (S,A)	2	1636, 1644 (E, NA), 1000 (S,NA&A)	3PLM	L_S	经验临界值
Yu & Cheng (2019)	NA	S&E	R	20, 40, 60, 80, 100, 120(S),19(E)	5(S), 4(E)	10000 (S), 6457 (E)	GRM	$W_{\max}, L_{\max}, S_{\max}, R_{\max}$	经验临界值
Yu & Cheng (2022)	NA	S&E	R	40,60,80(S)30(E)	2	1000 (S) 3000 (E)	2PLM	$W_{\max}, L_{\max}, S_{\max}$	经验临界值

注：A 表示自适应测验，NA 表示非自适应测验；S 表示模拟研究，E 表示实证研究；题目长度中的整数范围表示的是非定长 CAT 中的最小和最大长度；R 表示作答数据，RT 表示作答时间数据。

才能获得。

表 1 中虽然没有基于作答时间数据采用 CPA 方法进行检测的研究，但已有类似研究基于作答时间数据检测测验中的异常项目，例如 Choe 等(2018)使用序列分析方法分别基于作答数据、作答时间数据以及结合两种数据对泄露试题进行检测。该研究结果证明了在相同的 I 类错误率情况下，方法的检验力呈现：(1)仅基于作答时间数据的检验力要比仅使用作答数据的检验力高得多；(2)结合两种数据对泄露试题进行检测的方法有两种，其中一种方法的检验力略大于仅基于作答时间数据的检验力，第二种方法的检验力则远远小于仅基于作答时间数据。并且基于作答时间数据进行检测的探测点的识别延迟是 Choe 等(2018)所有方法中最小的。从 Choe 等(2018)的研究可以发现相比于作答数据，作答时间数据的确可以提供更多的测验信息，从而使检验力有实质性的提高。

如今作答时间数据的获取已经越来越容易，并且作答时间数据在检测考生的异常作答行为上比作答得分数据拥有先天的优势。比如当某位考生的作答模式为[1111101010]时仅从分数上不容易判断考生是否出现了加速作答，但是结合作答时间数据[57, 48, 51, 36, 42, 23, 18, 13, 7, 6]则更容易判断，因为加速作答的直接体现就是在作答时间上。因此基于作答时间数据检测异常作答行为是具有非常好的研究前景的。

3 基于 CPA 的统计量

从表 1 中可知，常用的 CPA 统计量大致有 4 种，分别是基于似然比检验的统计量($\nabla l_i, L_{\max}, L_S$)，基于 Wald 检验的统计量($W_{\max}, W_i^{(j)}$)，基于得分检验的统计量($R_S, S_{\max}$)和基于残差检验的统计量($R_{\max}$)。这几种统计量都是通过检验是否能拒绝虚无假设(考生的潜在特质没有发生显著的变化或考生的作答数据没有异常变化)来判断考生是否存在异常作答行为。本研究是基于作答时间数据检测异常作答，这里选用基于似然比检验和 Wald 检验的统计量。

本研究关注的异常作答行为是加速作答行为，因此在研究中使用单侧检验。考生出现加速作答行为后，考生的答题速度会增大。当变点 k 已知时，两个统计量的虚无假设为考生 i 在前 k 个题目上的速度参数等于后 $(J-k)$ 个题目上的速度参数，即 $\hat{\tau}_{i,k-} = \hat{\tau}_{i,k+}$。备择假设为考生 i 在前 k 个题目上的速度参数小于后 $(J-k)$ 个题目上的速度参数，即 $\hat{\tau}_{i,k-} < \hat{\tau}_{i,k+}$。下面对两种统计量基于作答时间数据时的形式进行介绍。

3.1 似然比检验

van der Linden (2006)提出的对数正态模型是应用最为广泛的作答时间模型，在很多实证研究中，它对作答时间数据的拟合都较好。本研究也使用该

模型，假设考生 i 在题目 j 上的作答时间 t_{ij} 服从以下密度函数：

$$f(t_{ij};\tau_i,\alpha_j,\beta_j)=\frac{\alpha_j}{t_{ij}\sqrt{2\pi}}exp\left\{-\frac{1}{2}[\alpha_j(\ln t_{ij}-(\beta_j-\tau_i))]^2\right\} \quad (10)$$

相当于

$$\ln(t_{ij})=\beta_j-\tau_i+\varepsilon_{ij},\varepsilon_{ij}\sim N(0,\alpha_j^{-2}) \quad (11)$$

其中 $\beta_j\in(-\infty,\infty)$ 为时间强度参数，β_j 值越大表明作答题目 j 需要花费的时间越长；$\tau_i\in(-\infty,\infty)$ 是考生 i 的速度参数，通常假定 τ 服从正态分布；α_j 为时间区分度参数，其作用类似于题目反应模型中的区分度参数。

基于作答时间模型(即公式 10)可得考生 i 的作答时间数据 t_i 的似然函数为：

$$L(\tau_i;t_i)=\prod_{j=1}^{n}\frac{\alpha_j}{t_{ij}\sqrt{2\pi}}\exp\left\{-\frac{1}{2}[\alpha_j(\ln t_{ij}-(\beta_j-\tau_i))]^2\right\} \quad (12)$$

其中 $t_i=(t_1,t_2,\cdots,t_n)$ 是考生 i 在所有题目上的作答时间数据。因此 $L(\tau_i;t_i)$ 的对数似然函数为

$$l(\tau_i;t_i)=lnL(\tau_i;t_i)=\sum_{j=1}^{n}\frac{\alpha_j}{t_{ij}\sqrt{2\pi}}-\frac{1}{2}\sum_{j=1}^{n}\{[\alpha_j(\ln t_{ij}-(\beta_j-\tau_i))]^2\} \quad (13)$$

似然比检验的公式与 Shao 等(2016)类似：

$$\Delta l_i^{(k)}=-2(l_i^{H_0}-l_i^{(k)}) \quad (14)$$

$l_i^{H_0}=l(\hat{\tau}_{i,0};t_i)$ 是通过使用考生在所有题目上的作答时间数据估计出的速度参数 $\hat{\tau}_{i,0}$ 计算得到的对数似然。假设考生的速度参数在题目 k 之后立即发生了突变，给定改变点 k 时，加速时的对数似然为：

$$l_i^{(k)}=l(\hat{\tau}_{i,k-};t_{i(k-)})+l(\hat{\tau}_{i,k+};t_{i(k+)}) \quad (15)$$

$\hat{\tau}_{i,k-}$ 是使用考生 i 在前 k 个题目上的作答时间数据 $(\hat{\tau}_{i(k-)}=t_{i1},t_{i2},\cdots,t_{ik})$ 估计出的速度参数，$\hat{\tau}_{i,k+}$ 是使用考生 i 在第 $k+1$ 至 J 个题目作答时间数据估计出的速度参数。由于在实际情况中，改变点的位置是未知的，因此将检验统计量设为所有可能的改变点位置上的 $\Delta l_i^{(k)}$ 的最大值：

$$\Delta l_{\max,i}=\max_{k=1,2,\cdots,(J-1)}\Delta l_i^{(k)} \quad (16)$$

当 $\Delta l_{\max,i}$ 在某个置信水平上超出可接受的范围时拒绝虚无假设，说明考生 i 的作答时间数据中出现了改变点。

3.2　Wald 检验

单侧 Wald 检验统计量的公式如下：

$$W_i^{(k)}=\frac{(\hat{\tau}_{i,k-}-\hat{\tau}_{i,k+})^2}{\frac{1}{I_{k-}(\hat{\tau}_{i,0})}+\frac{1}{I_{k+}(\hat{\tau}_{i,0})}} \quad (17)$$

其中 $I_{k-}(\hat{\tau}_{i,0})$ 和 $I_{k+}(\hat{\tau}_{i,0})$ 分别是基于考生 i 在前 k 个题目上和后 $(J-k)$ 个题目上的作答时间数据估计的 Fisher 信息量。当 k 未知时，检验统计量为

$$W_{\max,i}=\max_{k=1,2,\cdots,(J-1)}W_i^{(k)} \quad (18)$$

当 $W_{\max,i}$ 在某个置信水平上超出了可接受的范围时拒绝虚无假设，说明考生 i 的作答时间数据中出现了改变点。

本研究将 $l_i^{(k)}$ 和 $W_i^{(k)}$ 达到最大值的下一个项目作为加速点的估计值，即考生 i 从那个项目开始表现出加速作答行为。

3.3　CPA 统计量临界值的获取

和基于作答数据的似然比检验类似，基于作答时间数据的 $\Delta l_{\max,i}$ 也没有形成一个封闭的分布形态。参考表 1 中的信息，统计量临界值可通过置换分布、经验临界值和近似临界值获得。由于置换分布方法的计算量非常大，需要很长的时间获得临界值；而近似临界值比较适合改变点出现在测验中间位置(比如中间 70%的位置)的情况(Sinharay, 2016)。在实际情况中加速作答更容易出现在测验中晚期阶段，因为考生在测验的中晚期阶段更容易感受到时间的压力。因此本研究采用经验临界值。

当 k 已知时 $W_i^{(k)}$ 服从自由度为 1 的卡方分布，当 k 以及 $W_i^{(k)}$ 未知时，$W_{\max,i}$ 也不能形成一个封闭的分布形态。Sinharay (2016)指出，$W_{\max,i}$ 的渐进零分布和似然比统计量的渐进零分布是相同的。这里 Wald 检验统计量的临界值也采用经验临界值。具体过程如下。

参考 Worsley (1979)，通过公式 11，在测验长度为 40,60,80 的条件下随机生成 10000 个正常的作答时间模式。基于前面介绍的似然比统计量和 Wald 统计量的计算公式，分别得到 $\Delta l_{\max,i}$ 和 $W_{\max,i}$ 的 10000 个值；将它们按从大到小排序，得到它们第 500、第 100 和第 10 个最大值 $c_{0.05}$，$c_{0.01}$ 和 $c_{0.001}$，分别近似对应检验水平为 0.05, 0.01, 0.001 时的临界值。每种实验条件重复 100 次，取平均的 $c_{0.05}$，$c_{0.01}$ 和 $c_{0.001}$ 值作为后面实验中用到的经验临界值。

4　基于加速作答行为的作答时间模型

加速作答行为通常发生在有时间限制的考试中。当考试临近结束时，未完成作答的考生由于受

到时间因素的影响会倾向于提高自己的答题速度，出现加速作答。考生出现加速作答行为时，由于答题速度的增加，其作答时间会少于正常的答题时间。

为模拟考生加速作答行为下的作答时间，以往的研究提出了两种方法，第一种是将考生在加速作答行为下的作答时间设置为固定的几个水平，比如10 s, 20 s, 30 s (van der Linden & Guo, 2008)；第二种是在对数正态作答时间模型的参数 τ_i 上增加一个正数 L，表示加速作答对考生答题速度产生的影响。在 van der Linden 和 van Krimpen-Stoop (2003)的研究中，L 被设置为 0.375 和 0.750。这两种方式都将加速作答设置成了固定效应，即所有加速作答的考生都会出现相同的作答时间或受到相同大小的影响，这其实不太符合实际情况。

Yu 和 Cheng (2019)曾回顾了加速作答考生可能存在的两种潜在加速作答机制，即作答速度突变和作答速度逐渐改变，并使用两种模型来表示这两种作答机制，分别是混合模型(the hybrid model, HM)和逐渐变化模型(the graduate change model, GCM)。一方面，关于加速作答行为对于考生做题的影响机制，HM 假设考生出现加速作答时的作答速度会发生突变；而 GCM 认为考生在加速点之后的题目上的答对概率会逐渐下降；另一方面，有加速作答行为的考生，他们出现加速作答行为的位置是随机变量，即加速点各不相同。

在本研究中，我们拟采用更可能出现的作答速度“逐渐改变”的方式来模拟数据。Wollack 和 Cohen (2004)构建了基于作答数据的加速模型，该模型是模拟加速作答考生在题目上的正确概率发生“逐渐改变(下降)”，并且每位加速作答考生有其独特的加速模式。Goegebeur 等人(2008)进一步考察了该模型的参数估计。基于概率“逐渐改变”的三参数模型为：

$$P_{ij}^{*}=c_j+(1-c_j)\frac{exp[a_j(\theta_i-b_j)]}{1+exp[a_j(\theta_i-b_j)]}\times \min\left(1,\left[1-\left(\frac{j}{J}-\eta_i\right)\right]\right)^{\lambda_i} \quad (19)$$

其中，c_j 为猜测参数，$\frac{exp[a_j(\theta_i-b_j)]}{1+exp[a_j(\theta_i-b_j)]}$ 是常规的 2PLM，$\eta_i\,(0\leqslant\eta_i\leqslant1)$ 表示考生发生加速作答的位置，比如：$\eta_i=0.8$ 表示考生 i 在最后 20%的题目上出现加速作答。加入速度调节参数 λ_i 调节加速作答行为中正确作答概率下降的快慢，这个模型已经在很多研究中用来模拟加速作答数据(Shao et al., 2016; Suh et al., 2012)。

类似地，本研究中构建基于对数正态作答时间模型的作答时间逐渐下降模型，形式如下：

$$\ln(t_{ij})=(\beta_j-\tau_i+\varepsilon_{ij})\times \min\left(1,\left[1-\left(\frac{j}{J}-\eta_i\right)\right]\right)^{\lambda_i},\ \varepsilon_{ij}\sim N(0,\alpha_j^{-2}) \quad (20)$$

这个公式中的 η_i 和 λ_i 参数与上面公式中的参数含义相同。当测验没有进行到 η_i 所表示的阶段时 $\frac{j}{J}$ 将小于 η_i，故而 $\left[1-\left(\frac{j}{J}-\eta_i\right)\right]$ 的值会大于 1，$\min\left(1,\left[1-\left(\frac{j}{J}-\eta_i\right)\right]\right)^{\lambda_i}=1$，这说明考生的作答时间依旧使用对数正态作答时间模型模拟。当测验进行到 η_i 所表示的阶段时，即 $\frac{j}{J}>\eta_i$，则 $\min\left(1,\left[1-\left(\frac{j}{J}-\eta_i\right)\right]\right)^{\lambda_i}$ 小于 1；此时 $\ln(t_{ij})$ 的值将小于其正常的作答时间，表示考生在 η_i 所表示的阶段上出现了异常的加速作答行为。

5　模拟研究

为了完整的阐述 CPA 方法的使用过程以及评价 CPA 方法在作答时间数据上检测异常作答模式和加速作答行为上的表现，我们拟进行模拟研究和实证数据分析。通常来说，实证数据中的项目参数有可能是已知的(比如自适应测验系统)，也有可能是未知的。因此，我们考虑在已知项目参数和未知项目参数的条件下分别展开模拟研究。对于已知项目参数，实验中只需要估计考生的速度参数，采用 EAP 算法(Shao, 2016)；对于未知项目参数，项目参数基于全体考生的数据，采用 MCMC 算法(Fox et al., 2021)估计得到。

5.1　模拟研究设计

模拟研究中考生的数量固定为 1000，考虑 3 种测验长度分别 40, 60 和 80，它们的测验总时间分别设置为 60, 90 和 120 分钟(Shao, 2016)。考生中出现加速行为的比例为 10%，20%和 30%，分别表示加速作答行为的 3 种流行程度(低、中和高)。改变点位置 η_i 将从 4 种分布中生成，详细信息呈现在数据生成部分。当考试结束，考生还未完成所有的试题时测验直接终止，没有做完的题目的作答时间设置为 0，考生直接被标记为具有加速行为。模拟研究

共 $3\times3\times4\times2=72$ 种条件，每种条件重复 50 次(模拟条件见表 2)。模拟研究使用 R 程序完成。

表 2　模拟条件

因素	水平
测验长度	40, 60, 80
加速作答考生的比例	10%, 20%, 30%
改变点的位置参数 η_i	Median (0.6,0.7)× σ_η^2 (0.04,0.001)
项目参数	已知，未知

5.2　数据的生成

参考 Patton (2015)将题目区分度参数 a 和难度参数 b 设置为 $a\sim lnN(0,0.5), b\sim N(0,1)$。正常考生与具有加速行为考生的作答时间由公式(11)与(20)生成，其中时间区分度参数 α_j 服从均匀分布 $U(1.75,3.25)$；时间强度参数 β_j 和速度参数 τ_i 参考 Patton (2015)的设置：β_j 均值为 4，标准差为 1/3，β_j 与题目区分度参数 a 和难度参数 b 的相关系数设置为 0.3,0.5；$\tau_i\sim N(0,.25)$。对于具有加速作答行为的考生，我们采用与 Suh 等(2012)相同的做法来生成速度调节参数，即 $\lambda_i\sim \log N(3.912,1)$。改变点位置 η_i 按照 Shao 等(2016)的处理，即假定 η_i 服从 beta 分布，并且中值为 0.6 和 0.7，方差为 $\sigma_{\eta_i}^2=0.001$ 和 0.04，对应的 4 种 beta 分布具体形式为：beta (143.367, 95.689)，beta (2.970, 2.091)，beta (146.345, 62.910)和 beta (3.033, 1.490)。这样一来，改变点的分布如下图 1 所示。需要注意的是 η_i 是以百分比来反应加速作答的位置。对于测验长度为 40 的测验，$\eta_i=0.6$ 表示考生将从第 25 题开始加速作答。图 1 表明当 $\sigma_{\eta_i}^2=0.001$，生成的加速作答起点都接近于中值，而当 $\sigma_{\eta_i}^2=0.04$ 时，加速作答起点会更加分散，可能出现在测验的任何地方，甚至可能出现在接近测验结束的地方。在我们的研究中，考虑将接近测验中后期的地方作为加速作答的起点，主要原因有两个方面：首先是加速作答在通常情况下更容易出现在测验中后期；其次本研究想要考察当改变点不是在接近测验中间的位置时，近似临界值是否可以直接应用(Sinharay, 2016)。

两种统计量的临界值首先通过蒙特卡洛模拟生成，需要注意的是，在已知项目参数的实验条件，所有的计算结果都是基于项目参数真值得到；在未知项目参数时的实验条件，所有的计算结果都是基于项目参数的估计值得到。随后与 Sinharay (2016)中采用的近似临界值进行比较，选取合适的临界值用于加速作答行为的检验。

5.3　异常作答数据的检测过程

基于前文对 CPA 统计量的分析，研究拟使用似然比统计量与 Wald 统计量依次对每位考生的作答时间数据进行检测。大致过程如下：(1)计算每位

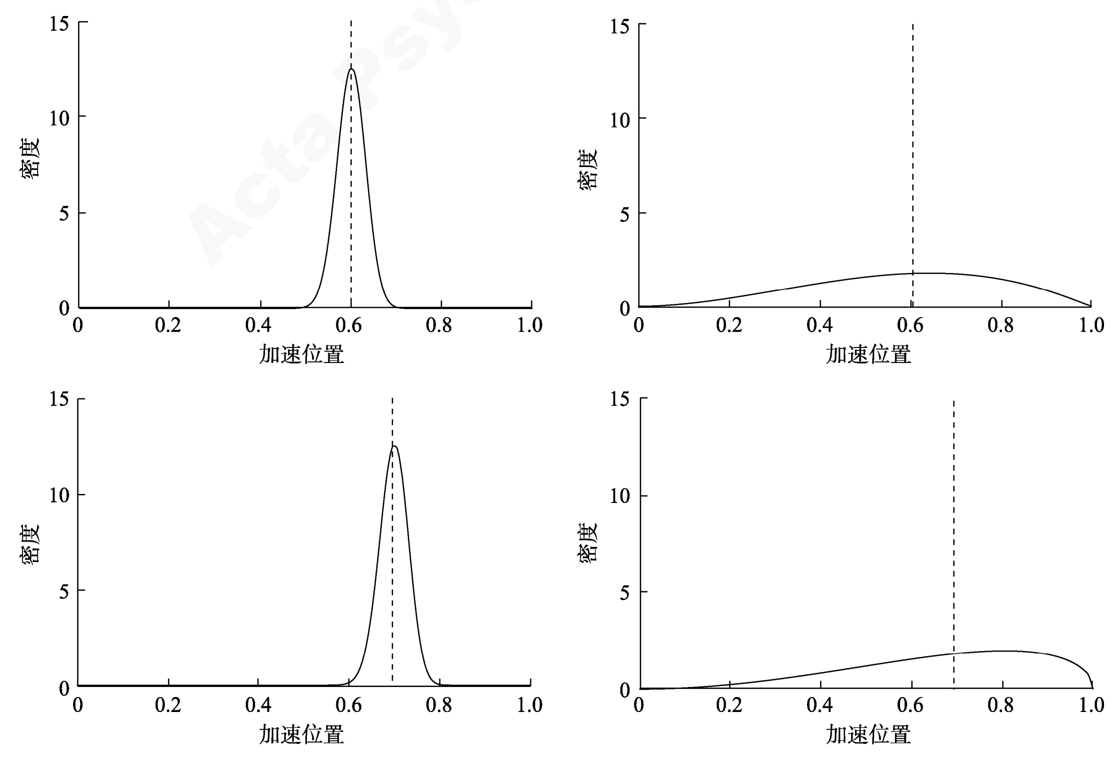


图 1　改变点的分布

考生在每道题上的似然比值,选取最大的似然比值作为似然比统计量的值, Wald 检验类似; (2)将两种方法的统计量与各自对应条件下的临界值进行比较，当统计量值超出临界值时，将考生的作答数据标记为异常作答数据; (3)当考生的作答数据标记为异常时，统计考生异常作答行为出现的位置; (4)用预定的评价指标对 CPA 方法的检测效果进行评价。

5.4 评价指标

使用 I 类错误率和检验力评价 CPA 方法的性能, I 类错误率和检验力的最终结果为每种条件下的均值。并计算在给定时间内未完成测验的学生比例(%*NF*)以及检测到的改变点位置与真实改变点位置之间绝对的延迟(absolute detection lag, ADL)指标的均值和标准差。I 类错误率、检验力和 *AL* 的计算公式分别如下

$$\text{I类错误率} = \frac{\text{错误标记加速考生的数量}}{\text{正常考生的总数}} \tag{21}$$

$$\text{检验力} = \frac{\text{正确标记加速考生的数量}}{\text{加速作答考生总数}} \tag{22}$$

$$ADL = \frac{\sum_{i=1}^{N} |\hat{p}_i - p_i|}{N} \tag{23}$$

其中, $\hat{p}_i$ 和 p_i 表示考生 i 由 CPA 方法探查到的改变点的位置和真实的改变点位置, N 是考生人数。

计算在给定时间内未完成测验的学生比例(%*NF*)是为了考察测验时间、测验长度等设置的是否合理，为测验设计提供一些有用的参考信息。

5.5 模拟研究结果

表 3 中分别给出了已知项目参数和未知项目参数时的临界值，呈现的是每种实验条件下临界值的平均值和标准差。由于似然比检验统计量和 Wald 检验统计量的经验临界值几乎相同，表 3 只给出了似然比检验统计量的临界值。一方面，从表 3 可以看出，未知项目参数对临界值的影响较小。另一方面，经验临界值随着检验水平有明显的变化，这与随测验长度的变化不同。随着测验长度的增加，临界值只有轻微的增加。经验临界值在 $\alpha = 0.05$ 和 0.01 时的方差比较小，表明临界值还是比较稳定的。对于 $\alpha = 0.001$ 时，方差比较大是可以解释的，因为临界值是基于 10,000 的样本，在分布的末端，统计量的值应该有更大的波动。

Sinharay (2016)中的表 1 (p.531)呈现了各置信水平下的近似临界值，当 $\alpha = 0.05$ 时列出的临界值从 8.45 到 9.84; 当 $\alpha = 0.01$ 时，临界值从 11.69 到 13.01。与 Sinharay (2016)中的近似临界值相比，表 3 中的数据虽然差异不是太大，但还是有些不同。正如前面所解释的，近似临界值可能更适合在长测验中使用，并且加速作答发生的位置容易出现在中后期的位置，而不是出现在测验早期。在我们的模拟中，测验长度相对都较短，并且加速作答的位置可以出现在测验的任何位置，这应该就是经验临界值与近似临界值出现较小差异的原因。在不同测验长度和重复实验上出现较稳定的经验临界值表明使用经验临界值是合适的。因此，根据表 3 所列的取值分别作为不同测验长度下 $\alpha_{0.05}$, $\alpha_{0.01}$ 和 $\alpha_{0.001}$ 的临界值。

表 3 似然比检验统计量对应的经验临界值的均值和标准差

项目参数	测验长度	$\alpha_{0.05}$	$\alpha_{0.01}$	$\alpha_{0.001}$
已知	40	8.068 (0.08)	11.214 (0.21)	15.702 (0.58)
	60	8.261 (0.07)	11.470 (0.20)	15.885 (0.58)
	80	8.352 (0.09)	11.732 (0.19)	16.247 (0.61)
未知	40	8.247 (0.12)	11.389 (0.30)	15.824 (0.65)
	60	8.353 (0.10)	11.517 (0.36)	15.889 (0.66)
	80	8.366(0.21)	11.798 (0.34)	16.456 (0.73)

表 4 和表 5 分别给出了已知和未知项目参数时各实验条件下的似然比统计量的检验力和 I 类错误率，从结果可以看出，两种统计量在已知项目参数时的表现要略好一些，但是它们在不同实验条件下表现的变化趋势比较一致。

5.5.1 已知项目参数的结果

所有条件下，除了在测验长度为 80，显著性水平为 0.001 时，其它条件下的 I 类错误率都只是稍微大于对应的显著性水平。与此同时，不论测验的长度，改变点的分布，以及受加速作答影响的被试比例，每种条件下的检验力都很高，很多情况下都接近于 1。比如：测验长度更长或受加速作答影响的题目比例越高时，检验力相对会更高。整的来说，与基于作答数据来说，基于作答时间数据对加速作答的检验力会高很多，比如，Shao 等人(2016)报告的基于作答得分数据检验加速行为的检验力在 0.60 到 0.90，而 Sinharay (2016)报告的检验力在多数条件下相对更低，Yu 和 Cheng (2022)报告的检验力也低于本研究中的检验力。这些都说明，基于作答时间检测异常作答行为更有优势。就 I 类错误率来说，每种条件下都能有很好的控制，都只是稍微大于对应的显著性水平。

表 4　模拟研究结果(已知项目参数)

测验长度	%	η_{median}	η_{var}	Power 0.05	Power 0.01	Power 0.001	Type-I-Error 0.05	Type-I-Error 0.01	Type-I-Error 0.001	%NF	ADL_{mean}	ADL_{SD}
40	10	0.6	0.001	0.98	0.98	0.98	0.053	0.012	0.0013	4.84	0.75	0.68
		0.7	0.001	0.97	0.97	0.97	0.053	0.012	0.0012	4.81	0.96	0.97
		0.6	0.04	0.96	0.96	0.95	0.055	0.013	0.0014	4.65	2.72	2.95
		0.7	0.04	0.94	0.94	0.93	0.051	0.012	0.0015	4.80	5.43	6.45
	20	0.6	0.001	0.98	0.97	0.96	0.054	0.011	0.0015	4.51	0.80	0.71
		0.7	0.001	0.96	0.96	0.95	0.052	0.011	0.0016	4.56	1.04	1.15
		0.6	0.04	0.96	0.96	0.95	0.052	0.012	0.0014	4.49	4.07	4.95
		0.7	0.04	0.94	0.93	0.93	0.053	0.013	0.0013	4.58	6.68	6.99
	30	0.6	0.001	0.96	0.96	0.95	0.056	0.012	0.0013	3.90	0.86	0.88
		0.7	0.001	0.94	0.94	0.93	0.052	0.013	0.0012	4.00	1.11	1.08
		0.6	0.04	0.95	0.95	0.94	0.054	0.011	0.0013	3.90	5.20	6.12
		0.7	0.04	0.93	0.93	0.93	0.053	0.012	0.0012	4.25	8.08	7.77
60	10	0.6	0.001	1.00	1.00	1.00	0.057	0.011	0.0016	6.78	1.05	1.62
		0.7	0.001	0.98	0.98	0.97	0.056	0.012	0.0015	6.88	1.34	1.39
		0.6	0.04	0.99	0.99	0.99	0.055	0.013	0.0014	6.95	5.67	7.55
		0.7	0.04	0.98	0.97	0.96	0.056	0.014	0.0013	7.24	7.86	9.48
	20	0.6	0.001	1.00	1.00	1.00	0.058	0.012	0.0014	6.65	1.38	1.75
		0.7	0.001	0.98	0.98	0.96	0.057	0.012	0.0016	6.57	1.64	1.82
		0.6	0.04	0.99	0.99	0.97	0.055	0.013	0.0017	6.64	7.05	7.67
		0.7	0.04	0.96	0.95	0.95	0.054	0.013	0.0014	6.82	9.12	9.91
	30	0.6	0.001	1.00	1.00	1.00	0.056	0.011	0.0015	6.19	1.88	1.83
		0.7	0.001	0.99	0.99	0.98	0.054	0.013	0.0013	6.08	2.09	1.92
		0.6	0.04	0.99	0.99	0.99	0.055	0.012	0.0013	5.99	9.21	9.46
		0.7	0.04	0.97	0.96	0.96	0.055	0.011	0.0018	6.25	10.78	10.71
80	10	0.6	0.001	1.00	1.00	1.00	0.067	0.015	0.0025	6.94	1.75	1.96
		0.7	0.001	1.00	1.00	1.00	0.076	0.017	0.0023	7.48	1.83	2.24
		0.6	0.04	1.00	1.00	1.00	0.071	0.016	0.0024	7.17	5.99	7.08
		0.7	0.04	1.00	1.00	0.99	0.074	0.015	0.0021	7.88	10.98	10.32
	20	0.6	0.001	1.00	1.00	1.00	0.072	0.016	0.0022	6.42	1.87	1.99
		0.7	0.001	1.00	1.00	1.00	0.073	0.017	0.0026	6.46	1.95	2.36
		0.6	0.04	1.00	1.00	1.00	0.075	0.015	0.0021	6.71	7.05	7.49
		0.7	0.04	1.00	1.00	0.98	0.074	0.016	0.0023	6.85	9.33	10.38
	30	0.6	0.001	1.00	1.00	1.00	0.073	0.016	0.0023	6.33	1.76	2.28
		0.7	0.001	1.00	1.00	1.00	0.074	0.015	0.0025	6.30	2.26	2.49
		0.6	0.04	0.99	0.99	0.99	0.077	0.017	0.0022	6.54	12.64	12.49
		0.7	0.04	0.98	0.99	0.98	0.073	0.016	0.0024	6.75	13.95	13.71

绝对延迟(*ADL*)指标的均值和标准差在 $\eta_{var} = 0.001$时很小。当 $\eta_{var} = 0.04$，并且测验长度为 80 时，这个延迟相对比较大，最大接近 14。如图 1 所示，当 η_{var} 较大，意味着改变点发生的位置可以在测验的任何位置，可以出现在测验的末期，这种情况下可能很难准确地检测到它真实发生的位置。已有的研究(Andrews, 1993; Hawkins et al., 2003)表明，基于 CPA 的方法更适合在中等长度的测验中检测异常发生的位置。比如：Andrews (1993)建议将搜索的范围限制为 $j = j_1, j_1+1, \cdots, n-j_1$，其中 j_1 大致等于 $0.15n$。换句话说，改变点发生的位置大致等于整个测验中间部分的 70%，这样可以保证提高检测改变点发生位置的准确度。

对于 $\%NF$，可以看出，在所有的条件下，在测

表 5 模拟研究结果(未知项目参数)

测验长度	%	η_{median}	η_{var}	Power			Type-I-Error			%NF	ADL_{mean}	ADL_{SD}
				0.05	0.01	0.001	0.05	0.01	0.001			
40	10	0.6	0.001	0.96	0.97	0.97	0.0568	0.0128	0.0013	4.99	0.76	0.71
		0.7	0.001	0.95	0.95	0.96	0.0570	0.0131	0.0014	4.83	0.97	1.03
		0.6	0.04	0.95	0.93	0.93	0.0638	0.0147	0.0015	4.72	2.73	2.98
		0.7	0.04	0.93	0.94	0.93	0.0513	0.0127	0.0015	4.83	5.55	6.49
	20	0.6	0.001	0.97	0.96	0.95	0.0625	0.0112	0.0015	4.71	0.86	0.76
		0.7	0.001	0.96	0.92	0.93	0.0530	0.0114	0.0017	4.57	1.04	1.18
		0.6	0.04	0.95	0.93	0.95	0.0531	0.0123	0.0015	4.54	4.11	4.95
		0.7	0.04	0.93	0.91	0.89	0.0584	0.0135	0.0013	4.60	6.76	6.99
	30	0.6	0.001	0.94	0.92	93	0.0588	0.0128	0.0014	3.96	0.94	0.90
		0.7	0.001	0.93	0.92	0.90	0.0626	0.0149	0.0014	4.04	1.14	1.11
		0.6	0.04	0.93	0.94	0.93	0.0655	0.0111	0.0013	3.97	5.24	6.13
		0.7	0.04	0.93	0.92	0.92	0.0589	0.0120	0.0013	4.26	8.15	7.80
60	10	0.6	0.001	0.99	0.99	0.98	0.0600	0.0116	0.0016	6.91	1.10	1.64
		0.7	0.001	0.97	0.94	0.95	0.0581	0.0136	0.0017	6.93	1.34	1.40
		0.6	0.04	0.97	0.96	0.97	0.0589	0.0143	0.0015	7.07	5.69	7.59
		0.7	0.04	0.96	0.96	0.94	0.0590	0.0146	0.0014	7.39	7.88	9.51
	20	0.6	0.001	0.98	0.99	0.98	0.0628	0.0124	0.0015	6.67	1.41	1.76
		0.7	0.001	0.97	0.96	0.95	0.0601	0.0120	0.0017	6.59	1.72	1.84
		0.6	0.04	0.96	0.99	0.92	0.0600	0.0131	0.0018	6.65	7.06	7.70
		0.7	0.04	0.94	0.92	0.95	0.0595	0.0145	0.0014	6.86	9.13	9.93
	30	0.6	0.001	0.98	0.99	0.97	0.0632	0.0114	0.0016	6.27	1.93	1.88
		0.7	0.001	0.99	0.99	0.97	0.0594	0.0136	0.0014	6.12	2.20	1.94
		0.6	0.04	0.97	0.98	0.99	0.0599	0.0135	0.0015	6.03	9.22	9.46
		0.7	0.04	0.95	0.92	0.95	0.0585	0.0114	0.0019	6.26	10.82	10.74
80	10	0.6	0.001	0.99	0.99	0.97	0.0715	0.0160	0.0026	6.95	1.79	1.96
		0.7	0.001	0.96	0.96	0.97	0.0817	0.0181	0.0024	7.49	1.84	2.27
		0.6	0.04	0.98	0.99	0.96	0.0777	0.0161	0.0025	7.19	6.00	7.10
		0.7	0.04	097	097	0.96	0.0742	0.0170	0.0023	7.97	11.07	10.39
	20	0.6	0.001	0.96	0.98	0.99	0.0776	0.0169	0.0022	6.46	1.89	2.01
		0.7	0.001	0.97	0.96	0.98	0.0745	0.0182	0.0027	6.49	1.98	2.39
		0.6	0.04	0.99	0.95	0.94	0.0758	0.0165	0.0023	6.78	7.12	7.50
		0.7	0.04	0.98	0.98	0.96	0.0788	0.0175	0.0024	7.01	9.36	10.42
	30	0.6	0.001	0.94	0.99	0.99	0.0757	0.0164	0.0024	6.55	1.78	2.30
		0.7	0.001	0.99	0.99	0.99	0.0745	0.0172	0.0026	6.40	2.32	2.53
		0.6	0.04	0.96	0.97	0.99	0.0798	0.0182	0.0023	6.64	12.67	12.52
		0.7	0.04	0.95	0.99	0.97	0.0787	0.0162	0.0027	6.78	14.00	13.71

验长度为 40 时，有 3.9%~5%的考生没有在预定的时间内完成测验；在测验长度为 60 时，分别有5.9%~7.3%的考生没有在预定的时间内完成测验。当受加速作答影响的被试比例达到 30%，长度为 80 的测验中有 5.9%~8%的考生没有在预定的时间内完成测验。这表明测验时间的设置是比较合理的，受加速作答影响的考生多数还是能在预定的时间内完成考试。本研究中，那些没有按时完成测验的考生会被标记成异常考生。表 4 中的检验力接近于 1 表明基于 CPA 的方法能够检测出那些不那么严重的加速作答考生，即那些存在加速作答行为，但是仍然在规定时间内完成测验的考生。

5.5.2 未知项目参数的结果

表 5 给出了未知项目参数时，似然比统计量在

不同条件下的表现。总体来说, 表 5 中的各实验条件下的统计检验力仍然很高, 最低值为 0.89, 一类错误率也得了很好的控制, 但是相对于表 4 中的结果略低, 说明未知项目参数时似然比统计量的表现仍然可靠, 只是略有下降。

表 5 中的结果显示出了和表 4 中的结果相同的趋势, 即随着测验长度的增加, 统计检验力略有上升, 比如 40 题时, 各条件下平均的检验力为 0.94 (α = 0.05), 0.93 (α= 0.01), 0.93 (α= 0.001), 60 题时, 各条件下平均的检验力为 0.97 (α= 0.05), 0.97 (α= 0.01), 0.96 (α= 0.001)。在 α 为 0.05 和 0.01 时, 各条件下的一类错误率接近显著性水平。当 α 为 0.001 时, 由于过于极端条件下得到的经验临界值导致各条件下的一类错误率相对较小。关于加速作答位置的估计, 可以看出测验长度的增加会导致位置估计值的绝对延迟(lag)变大, 测验长度从 40 增加到 80 时, 平均的 ADL_{mean} 从 3.19 增加到 5.99。异常位置的中值 η_{median} 和方差 η_{var} 也会影响其估计值, 尤其是较大的 η_{var} 会造成估计的位置有较大的延迟。当 η_{var} 为 0.04 时, 加速作答位置绝对延迟估计的平均值和方差分别为 7.92 和 8.44。

综合来看, 本研究中的实验条件与 Shao 等人(2016), Yu 和 Cheng (2022)相近, 虽然与表 4 和表 5 中的结果不能直接比较, 但是也还是有一定的指示作用。已知项目参数, 测验长度为 40 时, 基于得分数据的似然比检验和 Wald 检验统计量检验力在不同条件下的最小值和最大值分别为 0.50 和 0.94, 小于本研究中的 0.89 和 0.97。

6　实证数据分析

为了展示基于 CPA 的方法在实测数据中的使用, 我们将前面介绍的两种方法应用到实证数据, 该数据是某地区基础教育测量中的四年级的数学科目, 我们选了该试卷的一个题本。该题本的测验时间是 45 分钟, 包括 36,000 个考生在 30 题上的作答时间。所有的题目都是多项选择题, 在计算机上完成。我们首先对数据进行了整理, 将那些测试总时间过短(小于 5 分钟)的考生数据删除。同时为了考察 CPA 方法在检测更轻微加速作答的考生上的表现, 我们也删除了那些在测验末期题目上的作答时间为 0 的考生。最终有 33, 000 名考生的数据被保留下来, 我们从其中随机抽取 5000 名考生的作答时间数据进行分析。

基于这 5000 名考生的作答时间数据, 首先用对数作答时间模型进行拟合, 参数估计所用到的软件是 R 包 LNIRT (Fox et al., 2007, 2021)。得到各题目的参数 α_j 和 β_j 以及考生的速度参数 $\tau_{i,0}, \tau_{i,j-}$ 和 $\tau_{i,j+}$。5000 名考生的速度参数均值为 0 标准差为 0.267, 对应的直方图如图 2 所示, 被试速度呈负偏态分布。

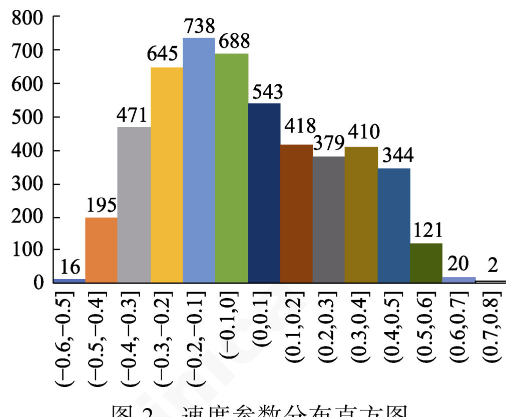


图 2　速度参数分布直方图

将得到的参数用于计算似然比检验统计量和 Wald 检验统计量的值。这两个统计量的结果非常接近, 因此, 我们这里只给出基于 Wald 检验统计量的结果。基于阈值 8.068, 在 5000 名考生中共有 675 位考生被检测出存在加速作答行为; 基于阈值 11.214, 共有 361 位考生被检测出存在加速作答行为; 基于阈值 15.702, 共有 271 位考生被检测出存在加速作答行为。图 3 显示了编号为 1034 考生的作答时间, 这个考生被标记为异常, 以及其期望作答时间、样本中异常考生的平均作答时间和样本中所有考生的平均作答时间。其中蓝色线表示的是 1034 号考生各题目的作答时间, 红色线表示的是测验中各题目的平均作答时间, 绿色线表示的是该考生的期望作答时间, 可以看出该考生的作答速度在前面 18 题是略高于平均速度的, 但是在这之后的题目上的作答速度是低于平均速度的。灰色线表示的是所有识别出的“异常”考生的平均作答时间。从图中可以看出 1034 号考生在最后 12 道题的作答时间都不超过 30 秒, 有几题在 10 秒左右, 和前面的 18 道题目相比作答时间有很大的下降。另外从图上可以看出, 无论是 1034 号考生还是全体考生, 测验后期题目的平均作答时间有下降的趋势, 说明考生接近考试结束时的题目倾向于花更少的时间完成, 并且所有 “异常” 考生的平均作答时间在测验后期下降的幅度更大。

所有被试在每道题目(除了最后的一道题)上的平均作答时间为 30~65 秒, 最后的一道题的平均作

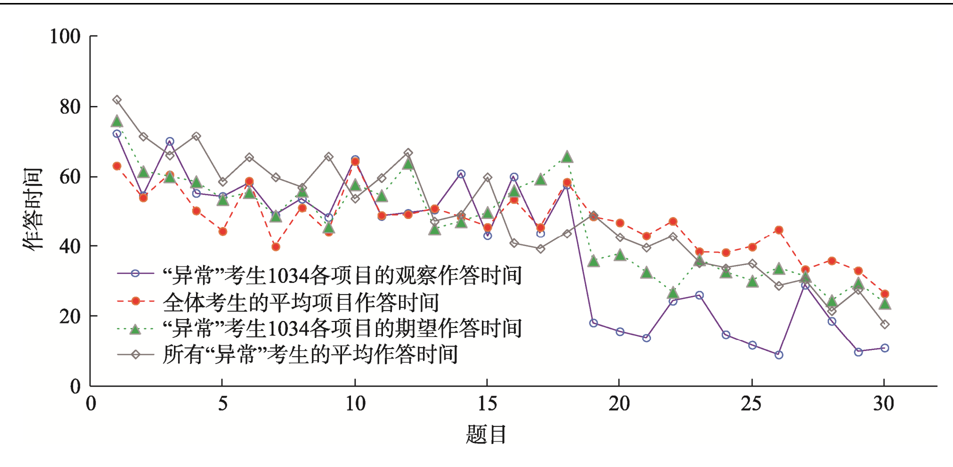


图 3 “异常”考生 1034 的各项目的作答时间、期望作答时间, 全体考生平均项目作答时间, 所有异常考生的平均项目作答时间

答时间接近 26 秒。从图中可以看出该考生在测验前期各题的作答时间是接近或大于平均作答时间的, 但是在 18 题之后, 各题的作答时间都小于平均作答时间, 与期望作答时间的差距明显变大了。这表明将编号为 1034 的考生标记为异常考生是合适的。

7 结论和讨论

本文基于 CPA 的两种检验统计量, 利用作答时间数据来检测具有加速作答行为的考生。通过模拟研究和实证数据分析对 CPA 方法的表现进行了评价, 结果显示两种检验统计量都具有非常高的检验力, 并且能够很好的控制 I 类错误率, 本研究表明基于作答时间数据在异常作答行为检测上具有很高的检验力。实际上, 基于得分数据和作答时间数据在检测异常作答行为时各有其特点, 将它们结合起来则有可能进一步对考生的异常作答行为的类型进行分析和探索。

本研究的结果进一步表明基于作答时间数据的 CPA 方法具有很好的研究前景, 而且它在实际应用中也具有可行性。首先, 本研究虽然采用对数作答时间模型来拟合作答时间数据, 但是基于 CPA 的方法可以扩展到其它作答时间模型比如 4 参数作答时间模型(Wang & Hanson, 2005) 或其它作答时间模型(Wang & Xu , 2015)下使用。其次, 无论项目参数是否已知, 本研究表 3 中的经验临界值在不同测验长度下相当接近, 表明可以在不同长度的测验使用相同的临界值, 这样使得方法的应用更加简单和方便, 比如, 当删除测验中的某些题或增加一些题目时, 没有必要再重新确定临界值。

将本研究中所涉及到的方法同时应用得分数据和作答时间数据上值得研究。一方面, 与得分数据相比, 基于作答时间的连续数据能够提供更丰富的信息, 作答时间数据在异常作答行为的检测上有优势; 另一方面, 得分数据有助于判断异常作答行为的类型(Wang et al,. 2018)。因此, 结合得分数据, 尤其是多级计分的得分数据(陈青 等, 2010; 程小杨 等, 2012)与作答时间数据检测异常作答行为和判断异常行为类型值得进一步探索和尝试。本研究表明, 不同测验长度可以使用相同的临界值, 因此, 本研究中的方法应该可以很容易推广到自适应测验 CAT 或多阶段自适应测验中(李佳, 丁树良, 2018; 熊建华 等, 2018)。需要注意的是, 加速作答只是一种较常见的异常作答行为, 本研究中的方法可以应用到检测其它类型的异常作答行为中, 比如应用到检测调查数据中的低作答动机考生等。

最后, 基于 CPA 的检测方法在多维测验中也是具有可行性和价值的。一方面现在多维测验逐渐增多, 例如在基于英语语言的数学测验中, 测验同时考察英语和数学两个维度的能力(张龙飞 等, 2020)。另一方法, 多维 RT 模型的数量也逐渐增多, 例如詹沛达等人(2020)开发了多维对数正态作答时间模型。其研究不但表明了在多维测验中, 潜在加工速度具有与潜在能力相匹配的多维结构, 还在模拟研究中实现了对被试的潜在加工速度的估计。这说明将 CPA 方法应用在多维测验中也是可以考虑和尝试的。

除了上文所说的, 本研究还有实际应用价值。开发不同的方法用于检测考生的异常作答行为是测验领域重要的质量控制解决方案。这个问题一直得不到很好的解决主要是由于题目参数和考生能力参数估计的不准确, 等值所带来的偏差以及对考生作答行为的不正确解释所造成的。比如, 一些终

结性测验的长度可能会很长，包含的内容也很多，这时需要综合进行考虑以确定合适的测验长度，让大部分的考生都有足够的时间完成测验(van der Linden, 2011; Patton, 2015; Patton et al.,2019)。在高风险测验中，考生在接近考试结束时间时会尽力完成所有的题目，对接近结束时间的题目采用快速猜测策略等。在这种情况下，没有完成测验中所有题目的考生比例可能会较少，因此需要有合适的方法来探查加速作答行为的流行程度。对于新开发的测验系统，建议能够记录下每位考生在每个题目上的作答时间，这可以为之后使用 CPA 方法来探测异常作答行为打下基础。

本研究虽然取得了一些有意义的结果，但还存在一些不足之处。首先，一般考试中通常记录的是考生的得分数据，以往也有研究基于得分数据检测加速作答，本研究中采用的是作答时间数据。当基于作答时间数据与基于得分数据的检测结果出现矛盾，仅从统计分析结果不容易判断哪种数据的检测结果是准确的时候，我们需要引入更多的信息(包括对测验内容的具体分析，其它统计量的分析，甚至是考场中的摄像记录和历史数据等)来谨慎地对这种数据做出综合评估(Wang et al., 2018)。其次，本研究应用的 CPA 方法需要假设改变点位置前后的作答概率模型是已知的。但是在实际应用中，改变点位置前后的概率结构可能是未知的。未来需要进一步探索不依赖于模型的改变点检测方法。另外，当 CPA 方法检测到了改变点时，我们只能推断作答数据发生改变的可能原因。例如，低作答动机和加速作答都可能会导致作答时间异常减少，本研究中的方法并不能对它们加以区分，也就是说，CPA 方法不能确定数据出现异常的原因。在这一点上我们需要结合其它的信息比如利用专家的领域知识来确定异常原因。未来还可结合作答数据和作答时间数据，进一步开发基于 CPA 的方法来检测不同的异常作答行为，充分发挥作答时间数据在检测异常作答行为上高检验力的优势。而且对于一些高风险测验，结合作答数据与作答时间数据共同做出推断会更加的合适。最后，目前使用 CPA 进行检测的研究大部分是基于大样本，未来可以尝试将 CPA 方法推广到小样本的情况中检测异常作答，观察 CPA 方法的检测效果。

参 考 文 献

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrical*, *61*(4), 821–856. https://doi.org/10.2307/2951764.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (second edition). Taylor & Francis Group. http://ebookcentral.proquest.com/lib/pqopenlayer/detail.action?docID=5378595.

Bejar, I. I. (1985). *Test speededness under number-right scoring: An analysis of the test of English as a foreign language*. (Report No. ETS-RR-85-11). Princeton, NJ: Educational Testing Services. https://doi.org/10.1002/j.2330-8516.1985.tb00096.x.

Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, *40*(2), 83–97. https://doi.org/10.1177/0146621615603327.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), 331–348. https://doi.org/10.1111/j.1745-3984.2002.tb01146.x.

Bradlow, E. T., Weiss, R. E., & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical Association*, *93*(443), 910–919. https://doi.org/10.1080/01621459.1998.1047374.

Chen, Q., Ding, S. L., Zhu, L. Y., & Xu, Z. Y. (2010). Three-parameter graded response model and its parameter estimation. *Journal of Jiangxi Normal University (Natural Science)*, *34*(2), 117–122.

[陈青, 丁树良, 朱隆尹, 许志勇. (2010). 3 参数等级反应模型及其参数估计. *江西师范大学学报(自然科学版)*, *34*(2), 117–122.]

Cheng, X. Y., Ding, S. L., Zhu, L. Y., & Wu, H. F. (2012). The stratified item selection strategy with maximal information under graded response model. *Journal of Jiangxi Normal University (Natural Science)*, *36*(5), 446–451.

[程小杨, 丁树良, 朱隆尹, 巫华芳. (2012). 等级评分模型下的最大信息量分层选题策略. *江西师范大学学报(自然科学版)*, *36*(5), 446–451.]

Choe, E. M., Zhang, J., & Chang, H.-H. (2018). Sequential detection of compromised items using response times in computerized adaptive testing. *Psychometrika*, *83*(3), 650–673. https://doi.org/10.1007/s11336-017-9596-3.

Evans, F. R., & Reilly, R. R. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement*, *9*(2), 123–131. https://doi.org/10.1111/j.1745-3984.1972.tb00767.x.

Fox, J.-P., Entink, R. K., & van der Linden, W. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software*, *20*(7), 1–14. https://doi.org/10.18637/jss.v020.i07.

Fox, J.-P., Klotzke, K., & Simsek, A. S. (2021). LNIRT: An R package for joint modeling of response accuracy and times. *ArXiv:2106.10144 [Stat]*. http://arxiv.org/abs/2106.10144.

Goegebeur, Y., de Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika*, *73*(1), 65–87. https://doi.org/10.1007/s11336-007-9031-2.

Guo, J., Tay, L., & Drasgow, F. (2009). Conspiracies and test compromise: An evaluation of the resistance of test systems to small-scale cheating. *International Journal of Testing*, *9*(4), 283–309. https://doi.org/10.1080/15305050903351901.

Guo, X. J., & Luo, Z. S. (2019). A psychometric model for speed-accuracy tradeoff and application. *Psychological Exploation, 39*(5), 451–460.
[郭小军, 罗照盛. (2019). 基于速度与准确率权衡的心理测量学模型及应用. *心理学探新, 39*(5), 451–460.]
Hawkins, D. M., Qiu, P., & Kang, C. W. (2003). The changepoint model for statistical process control. *Journal of Quality Technology*, *35*(4), 355-366. https://doi.org/10.1080/00224065.2003.11980233.
Li, J., & Ding, S. (2018). The several stratified methods of CAT in the presence of calibration error on GRM. *Journal of Jiangxi Normal University (Natural Science), 42*(4), 374–378.
[李佳, 丁树良. (2018). 基于 GRM 模型的 CAT 分层方法在校准误差中的应用研究. *江西师范大学学报(自然科学版)*, *42*(4), 374–378.]
Luo F., Wang X., Xu Y., & Feng W. (2020). Research progress of cheating detection technology in examinations: Detection of group cheating. *China Examinations*, (11), 37–41.
[骆方, 王欣夷, 徐永泽, 封慰. (2020). 考试作弊甄别技术的研究进展：团体作弊的甄别. *中国考试,* (11), 37–41.]
Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*(6), 426–451. https://doi.org/10.3102/1076998614559412.
McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, *27*(2), 121–137. https://doi.org/10.1177/0146621602250534.
Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*(3), 200–219. https://doi.org/10.1111/j.1745-3984.1994.tb00443.x.
Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, *42*(3/4), 523–527. https://doi.org/10.2307/2333401.
Pan, Y., & Wollack, J. A. (2021). An unsupervised-learning-based approach to compromised items detection. *Journal of Educational Measurement*, *58*(3), 413–433. https://doi.org/10.1111/jedm.12299.
Patton, J., Cheng, Y., Hong, M., & Diao, Q. (2019). Detection and treatment of careless responses to improve item parameter estimation. *Journal of Educational and Behavioral Statistics, 44*(3), 309–341. https://doi.org/10.3102/1076998618825116.
Patton, J. M. (2015). *Some consequences of response time model misspecification in educational measurement* (Unpublished doctoral dissertation). University of Notre Dame.
Shao, C. (2016). *Aberrant response detection using change-point analysis* (Unpublished doctoral dissertation). University of Notre Dame. https://curate.nd.edu/show/5425k932c5j.
Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika*, *81*(4), 1118–1141. https://doi.org/10.1007/s11336-015-9476-7.
Sinharay, S. (2016). Person fit analysis in computerized adaptive testing using tests for a change point. *Journal of Educational and Behavioral Statistics*, *41*(5), 521–549. https://doi.org/10.3102/1076998616658331.
Sinharay, S. (2017a). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*, *42*(1), 46–68. https://doi.org/10.3102/1076998616673872.
Sinharay, S. (2017b). Some remarks on applications of tests for detecting a change point to psychometric problems. *Psychometrika*, *82*(4), 1149–1161. https://doi.org/10.1007/s11336-016-9531-z.
Sinharay, S. (2017c). Which statistic should be used to detect item pre-knowledge when the set of compromised items is known? *Applied Psychological Measurement*, *41*(6), 403–421. https://doi.org/10.1177/0146621617698453.
Stefan, Z., Dietrich, K., & Wolfgang, H. (2016). Are exam questions known in advance? Using local dependence to detect cheating. *PLOS ONE*, *11*(12), e0167545. https://doi.org/10.1371/journal.pone.0167545.
Suh, Y., Cho, S.-J., & Wollack, J. A. (2012). A comparison of item calibration procedures in the presence of test speededness. *Journal of Educational Measurement*, *49*(3), 285–311. https://doi.org/10.1111/j.1745-3984.2012.00176.x.
van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204. https://doi.org/10.3102/10769986031002181.
van der Linden, W. J. (2011). Test design and speededness. *Journal of Educational Measurement*, *48*(1), 44–60. https://doi.org/10.1111/j.1745-3984.2010.00130.x.
van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365–384. https://doi.org/10.1007/s11336-007-9046-8.
van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*(2), 251–265. https://doi.org/10.1007/BF02294800.
Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477. https://doi.org/10.1111/bmsp.12054.
Wang, C., Xu, G., & Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, *83*(1), 223–254. https://doi.org/10.1007/s11336-016-9525-x.
Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*(5), 323–339. https://doi.org/10.1177/0146621605275984.
Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2.
Wollack, J. A., & Cohen, A. S. (2004). *A model for simulating speeded test data*. Paper presented at annual meeting of the American Educational Research Association, San Diego, CA. https://testing.wisc.edu/research papers/AERA 2004 (Wollack & Cohen).pdf.
Worsley, K. J. (1979). On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association*, *74*(366a), 365–367. https://doi.org/10.1080/01621459.1979.10482519.
Xiong, J., Luo, H., Wang, X., & Ding, S. (2018). The online calibration based on graded response model. *Journal of Jiangxi Normal University (Natural Science), 42*(1), 62–66.
[熊建华, 罗慧, 王晓庆, 丁树良. (2018). 基于 GRM 的在线校准研究. *江西师范大学学报(自然科学版)*, *42*(1), 62–66.]
Yu, X. F., & Cheng, Y. (2019). A change-point analysis

procedure based on weighted residuals to detect back random responding. *Psychological Methods*, *24*(5), 658–674. https://doi.org/10.1037/met0000212.

Yu, X. F., & Cheng, Y. (2022). A comprehensive review and comparison of CUSUM and change-point-analysis methods to detect test speededness. *Multivariate Behavioral Research*, *57*(1), 112–133. https://doi.org/10.1080/00273171.2020.1809981.

Zhan, P. D. (2019). Joint modeling for response times and response accuracy in computer-based multidimensional assessments. *Journal of Psychological Science*, *42*(1), 170–178.

[詹沛达. (2019). 计算机化多维测验中作答时间和作答精度数据的联合分析. *心理科学, 42*(1), 170–178.]

Zhan, P. D., Hong, J., & Man, K. W. (2020). The multidimensional log-normal response time model: An exploration of the multidimensionality of latent processing speed. *Acta Psychologica Sinica*, *52*(9), 1132–1142.

[詹沛达, Hong Jiao, Kaiwen Man. (2020). 多维对数正态作答时间模型：对潜在加工速度多维性的探究. *心理学报, 52*(9), 1132–1142.]

Zhang, J. (2014). A sequential procedure for detecting compromised items in the item pool of a CAT system. *Applied Psychological Measurement*, *38*(2), 87–104. https://doi.org/10.1177/0146621613510062

Zhang, L., Wang, X., Cai, Y., & Tu, D. (2020). Change point analysis: A new method to detect aberrant responses in psychological and educational testing. *Advances in Psychological Science*, *28*(9), 1462–1477.

[张龙飞, 王晓雯, 蔡艳, 涂冬波. (2020). 心理与教育测验中异常反应侦查新技术：变点分析法. *心理科学进展*, *28*(9), 1462–1477.]

Exploration of change point analysis in detecting speededness based on response time data with known/unknown item parameters

ZHONG Xiaoyuan[1], YU Xiaofeng[1], MIAO Ying[1], QIN Chunying[2], PENG Yafeng[1], TONG Hao[1]

([1] *School of Psychology, Jiangxi Normal University, Nanchang* 330022, *China*)

([2] *School of Mathematics and Information Science, Nanchang Normal University, Nanchang* 330032, *China*)

Abstract

In recent years, response time has received a rapidly growing amount of attention in psychometric research, likely due to the increasing availability of (item-level) response time data through computer-based testing and online survey data collection. Compared to the conventional item response data that are often dichotomous or polytomous, the response time is continuous and can provide much more information. Aberrant response behaviors are frequently encountered during testing. It could cause various negative effects. Change point analysis (CPA) is a well-established statistical process control method to detect changes in a sequence, and it has provided testing professionals a new lens through to understand test-taking behavior at both the examinee and item levels.

In this paper, we took test speededness as an example to illustrate how the CPA method can be used to detect aberrant behavior using item response time data. Response time under speededness was simulated using the gradual-change log-normal model for response time. Two CPA-based test statistics, the Likelihood Ratio Test and Wald Test, were used to detect aberrant response behaviors. The critical values were obtained through Monte Carlo simulations and compared with the approximate critical values in a previous study. Based on the chosen critical values, we examined the performance of the likelihood ratio test and Wald test in detecting speeded responses, specifically in terms of power and empirical Type-I error.

On the one hand, the critical values are almost identical for Wald and the likelihood ratio test. They vary substantially at different nominal α levels, but do not differ much across different test lengths. On the other hand, compared to approximate critical values, the critical values are not too far away from them but are different. That may be because the approximate critical values are suitable for situations where the change point appears in the middle of the test. Results indicate that the proposed method is much more powerful based on the critical values than conventional methods that use item response data. The power was close to 1 for most of the conditions while keeping the type-I error rate well-controlled. Real data analysis also demonstrates the performance of the method.

This study uses CPA with response time data and offers a very promising approach to detecting aberrant response behavior. Through the simulation study, we demonstrated that it is possible to use fixed critical values

in different test lengths, which makes the application of the method straightforward. It also means that it is unnecessary to reconduct the simulation to update critical values when small changes occur in the test. CPA is very flexible. This study assumed that the log-normal model fits the response time data, but the method is not bounded by that assumption.

Key words change point analysis, aberrant response behaviors, response time, test speededness, statistical process control